心理科学进展　2020, Vol. 28, No. 11, 1901–1910
Advances in Psychological Science

https://dx.doi.org/10.3724/SP.J.1042.2020.01901

利他性惩罚的动机*

陈思静　杨莎莎

(浙江科技学院经济与管理学院，杭州 310023)

摘　要　社会规范的维系离不开对违规者实施的利他性惩罚，然而，从个体心理层面来看，利他性惩罚的动机却并不是全然利他的。除了积极维护公平原则以外，追求良好声誉、规避潜在损失或消除负性情绪也在不同程度上驱动了利他性惩罚。此外，对惩罚成本数量和成本形式的敏感也表明基于成本–收益原则的策略性动机在驱动利他性惩罚中占据一席之地。进一步探索在利他性惩罚实施过程中不同动机之间的相互作用是未来的重要研究方向。

关键词　利他性惩罚，惩罚动机，公平原则，声誉，惩罚成本

分类号　B849: C91

1　引言

Thaler (1988)最早在最后通牒博弈中注意到被试在遭受不公平对待时，愿意牺牲自己的利益来惩罚违反了公平规范的分配者。Fehr 和 Gächter (2002)进一步发现，类似现象也会发生在第三方身上，即利益未受直接影响的第三方同样愿意付出成本去惩罚违规者。上述两种惩罚行为具有若干共同特征：1)给违规者造成了损失; 2)惩罚者需付出一定成本; 3)惩罚行为在一定程度上维护了相应的社会规范(李佳 等, 2012)。因此，目前学者通常使用利他性惩罚(altruistic punishment)这一术语来指称上述两类惩罚(谢东杰，苏彦捷，2019; Rodrigues et al., 2020)。

研究者在使用利他性惩罚这一术语时，“利他”一词的含义多是基于对惩罚结果的考量，即惩罚者通过牺牲自身利益维护了社会规范(Enge et al., 2017; Fehr & Gächter, 2002)。这种思路通常从远因角度(ultimate perspective)来理解利他性惩罚的演化机制，即利他性惩罚的演化背景、选择优势及其结果，来自生物学、经济学或博弈论等领域的研究者在这方面做出了巨大贡献。然而，正如 Henriques (2008)指出，如果我们希望建立关于某一社会现象的完整知识体系，那么心理学就应该在不同社会科学之间扮演一个联通与整合的角色。Rand 和 Nowak (2013)注意到，目前的主流演化模型倾向于将个体简化为不具备动机的行动者(agent)，这无疑会削弱我们对相关现象的理解，因为个体在行动时往往具有多种多样甚至相互冲突的动机(Svensson, 2020)，因此，如何从心理学角度评估不同动机对行为的影响有助于进一步理解利他性惩罚。对动机的忽视也导致了这一术语的含义在一定程度上变得模糊不清并阻碍了学者间的有效交流：如 Fehr 和 Gächter (2002)更多地在结果层面上使用这一概念，即利他性惩罚从结果的角度来说有利于社会规范的维系以及合作水平的提升，而 Rodrigues 等(2020)则侧重在动机层面使用利他一词，即利他意味着个体必须在主观上持有提升他人福祉的意图。而从动机角度出发，“利他性惩罚究竟在多大程度上是利他的”是目前学界争论的焦点之一(Pedersen et al., 2018)。为了更好地理解利他性惩罚，我们需要更深入考察其背后的动机。

Elster (2006)区分了两种利他行为模式：第一种模式中，个体是否做出利他行为主要取决于他人如何行动，而在第二种模式中，则更多地取决于自身行为是否会被他人观察到。单次匿名博弈

收稿日期：2020-03-17
* 国家自然科学基金项目(71701185)、浙江省软科学项目(2020C35020)资助。
通信作者：陈思静, E-mail: beepoison@163.com

中的利他性惩罚似乎更接近第一种模式，即惩罚主要取决于违规行为是否发生，而惩罚本身是否能被他人观察到并不重要。此时，利他性惩罚主要由针对违规者的愤怒等情绪所驱动，以成本–收益为特征的策略性考虑(strategic consideration)在这种情况下并不重要，而个体的决策机制被编码在直觉神经过程中，因此它常常是自发的并易受情感因素的影响(Camerer et al., 2005)。相反，在多次或/和非匿名博弈中，Elster (2006)所定义的第二种模式似乎发挥了更大作用。在这种情况下，惩罚动机往往具有功利含义，如威慑潜在的违规者(McCullough et al., 2013)、避免负面的道德评价(Gardner, 2019)、展现自身的优良特质(Jordan, Hoffman et al., 2016)或提升自己的声誉(Jordan & Rand, 2020)等。

Rodrigues 等(2020)指出，利他行为背后往往存在多种动机，对上述文献的简单回顾在一定程度上证实了这一点。这些研究表明，利他性惩罚的动机并不是单一的，换言之，它们可能由一系列相互制约的因素构成，从而形成一个复杂的系统。这就需要我们从既有的动机理论出发进一步梳理和分析利他性惩罚背后的动机及其影响因素。这一方面有助于厘清这一术语的准确含义，从而方便不同研究者之间的交流，另一方面也可为利他性惩罚的研究提供一个从个体心理层面出发的近因视角(proximate perspective)，即引导个体在特定情境下做出利他性惩罚的心理机制，从而为现有文献中占主导地位的远因视角做一个有益补充。同时，从实践角度来说，探明利他性惩罚背后的动机有助于更有针对性地制定政策，以有效地激励人们采取利他性惩罚，从而促进社会规范的维系与合作水平的提升。

2 理论框架

享乐主义原则(hedonic principle)长期以来在动机领域占有主导地位，该原则强调了个体在决策时趋乐避苦的动机(姚琦，乐国安，2009)。Higgins (1997)在此基础上，提出了调节焦点理论(regulatory focus theory)，进一步将个体的动机系统分为促进焦点(promotion focus)和预防焦点(prevention focus)：促进焦点和个体趋近积极目标状态以及实现增长与进步有关；而预防焦点则旨在规避消极状态和维护安全(Higgins, 1997; Whitson et al., 2019; Winterheld & Simpson, 2016)。研究者发现不同的调节焦点对个体的行为策略具有显著不同的影响(Crowe & Higgins, 1997)。就利他性惩罚而言，其动机也可以在一定程度上通过调节焦点理论来解释，例如，促进焦点占据主导地位时，利他性惩罚更多地表现为追求公平(Fehr & Schurtenberger, 2018)或积极的声誉(Jordan & Rand, 2020)；而当预防焦点更为突出时，个体的利他性惩罚则主要为了避免损失(Pedersen et al., 2018)。

另一个具有较大影响力的动机理论是由 Deci 和 Ryan (1985)所提出的自我决定理论(self-determination theory)。该理论区分了两种类型的动机：自主动机(autonomous motivation)和受控动机(controlled motivation)。自主动机的个体主要出于自己的主观意愿或信念而行动；相反，受控动机指的是个体出于外部压力(如报酬或群体规则)而实施某行为的动机。后续研究表明，自主–受控动机体系表现在广泛的行为模式中并具有高度的跨文化一致性(Deci et al., 2017; Roth et al., 2007)。这里需要强调的是，不同于简单的内外动机二分法，自主–受控动机系统体现了动机通过自我调节从外部转向内部的动态连续过程(胡小勇，郭永玉, 2009)。这种连续体的观点意味着自主–受控这两类动机并非截然可分，而是往往以不同比例交织在一起，正如我们在利他性惩罚动机中所观察到的一样。

基于上述两个理论，本文将利他性惩罚的动机划分为以下 4 类：1)自主–促进：追求公平；2)自主–预防：减少负性情绪；3)受控–促进：追求名誉；4)受控–预防：避免损失。图 1 展示了 4 种类型的惩罚动机在两个维度上的分布。此外，本文还进一步梳理了在个体惩罚决策过程中影响上述 4 类动机的不同因素，如社会距离(Baumgartner et al., 2012)、社会地位(van Prooijen & Lam, 2007)和惩罚成本(殷西乐 等, 2019)等。需要指出的是，上述分类在一定程度上是为了方便论述，而在现实中不同类别之间可能并不存在一个明确的分界点。

3 自主动机

3.1 促进焦点及其影响因素

Falk 等(2005)指出，针对违规者的利他性惩罚是一种非策略性惩罚(non-strategic sanction),

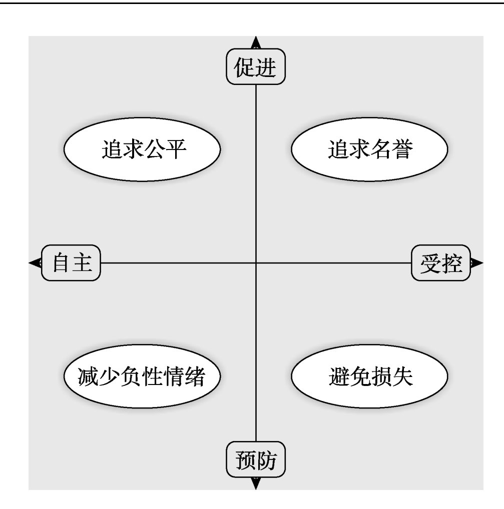


图 1　利他性惩罚的动机分布

主要由公平原则驱动, 因为在实验中外在奖惩条件的变化并没有显著改变惩罚者的行为模式, 这意味着个体实施利他性惩罚在很大程度上是为了维护自身的某种内在信念, 而非追求外在酬报。这种观点同样反映在了有关公正世界信念理论(belief in a just world)的研究中(Lerner, 1965)。该理论认为人们需要相信自己所生活的世界是公平的, 即在这样一个世界里善有善报恶有恶报。为了积极维护该信念, 个体愿意付出一定代价来惩罚破坏公正世界的违规者(Strelan et al., 2017)。因而, 这类惩罚动机在一定程度上体现了惩罚行为的自主−促进焦点。此外, 分配结果的不公平程度会显著影响被试的惩罚行为。具体而言, 分配结果越不公平, 做出惩罚行为的被试比例就越高(Fehr & Schmidt, 1999)。来自神经科学的研究同样说明了这一点: 前脑岛对不公平程度较为敏感, 随着不公平程度的增加, 该脑区的激活程度也随之增强, 而前脑岛激活程度越高的个体实施惩罚的概率也越大(Tabibnia et al., 2008)。特别能说明问题的是, 无论不公平是指向自己还是他人, 左侧前脑岛都十分敏感(Corradi-Dell'Acqua et al., 2013)。上述结果在一定程度上表明人们具有维护公平规范的内在动机, 而且当违规行为偏离公平规范的程度越高, 人们的惩罚动机也就越强烈(Fehr & Schurtenberger, 2018)。正如陈思静等(2015)指出, 利他性惩罚本身即是一个社会规范(social norm)激活的结果, 而分配结果的不公平程度能影响公平规范的激活水平, 从而影响个体的惩罚行为。这意味着利他性惩罚与维护公平规范之间的确存在着紧密关系。

需要说明的是, 利他性惩罚中对公平的追求受到个体和情境因素的影响。Fehr 和 Schmidt (1999)注意到面对不公现象, 个体在接受程度方面存在差异。此外, Cui 等(2019)也注意到, 面对不同公平程度的分配方案时, 高利他个体和低利他个体在晚期正成分(LPC)上的表现具有显著差异。通常来说, 越是注重公平的个体, 在面对不公平现象时就越容易实施利他性惩罚(Johnson et al., 2009)。陈思静和马剑虹(2011)的研究也得出了相似结论: 社会责任感高的个体在面对不公平现象时对违规者实施了更多的惩罚。

就情境因素而言, 利他性惩罚中的公平动机受到损失/收益框架的影响。在最后通牒博弈中, 相较于不公平的收益分配方案, 个体更倾向于拒绝不公平的损失分配方案, 这表明个体在损失框架下对公平的感知程度更高(Zhou & Wu, 2011)。此外, 社会距离(social distance)同样值得我们关注, 它反映了个体之间的相似性或亲近度, 是影响个体惩罚决策的重要因素(Guo et al., 2019; Vekaria et al., 2017)。徐杰等(2017)的研究表明, 社会距离可能通过调节个体的公平感知来影响其行为决策, 当看到社会距离较近的个体遭遇不公平待遇时, 第三方更加倾向于对违规者做出惩罚; 此外, 在攻击范式下, 目睹陌生人受到侮辱的第三方尽管感到愤怒, 但惩罚行为显著少于被侮辱者是亲友的情况(Pedersen et al., 2018)。总体而言, 人们倾向于向群体内部的受害者提供更多的补偿, 并且对来自群体外部的违规者施加更多的惩罚(Guala & Filippin, 2017)。

3.2　预防焦点及其影响因素

和自主−促进焦点不同, 自主−预防焦点意味着个体在进行利他性惩罚时主要是为了规避某种消极的内在状态。研究者发现, 当个体自己遭遇或目睹他人遭遇不公对待时, 就会产生一系列负性情绪, 如愤怒(anger) (Jordan, McAuliffe et al., 2016)、内疚(guilt) (Nelissen & Zeelenberg, 2009)或烦躁(upset) (Dimitroff et al., 2020)。这些情绪可以较好地正向预测个体进行利他性惩罚的概率。Harlé 和 Sanfey (2007)的研究显示, 如果实验者在博弈前通过影片唤起了被试的负性情绪, 他们更有可能惩罚不公平的分配方案; 而唤起正性情绪的被试, 其行为与对照组并无显著差异, 这表明

被试可能通过惩罚违规者来疏泄自己的负性情绪。此外, Jordan, McAuliffe 等(2016)发现, 负性情绪是驱使第三方维护社会规范的重要动机。对于上述结果, van Doorn 等(2014)的一个解释是愤怒等负性情绪是对不公平行为的反应, 它引发了个体恢复公平的冲动。因此, 在这种情况下, 个体进行利他性惩罚在一定程度上是为了避免或消除负性情绪。事实上, Xiao 和 Houser (2005)的研究证实了这一点, 他们发现, 如果被试在进行惩罚决策前其负性情绪得到了表达, 那么惩罚概率就会显著降低。来自神经科学的研究也证实了这一点, 如吴燕和罗跃嘉(2011)注意到, 当人们看到违规行为未受到惩罚时会产生心理不适, 而这种不适会随着惩罚的实施而降低。这也从侧面验证了 Hu 等(2016)的结论:利他行为有助于个体获得正性情绪。需要说明的是, 尽管为了论述方便, 我们将利他性惩罚的自主动机划分成了促进和预防两种焦点, 但在实际生活中, 这两种焦点更可能是一体两面: 目睹社会规范遭到破坏引发个体产生了恢复公平的冲动和愤怒等负性情绪, 而利他性惩罚则在一定程度上满足了个体的上述冲动并平息了负性情绪。

和惩罚行为的自主-促进焦点一样, 预防焦点同样受到一系列因素的影响。关于旁观者效应(bystander effect)的研究表明, 他人的存在会影响个体的情绪启动(Voelpel et al., 2008), 减弱个体责任感(Feng et al., 2016), 进而抑制亲社会行为(Panchanathan et al., 2013)。Nelissen 和 Zeelenberg (2009)证实了这种效应, 他们将被试分为有旁观者和无旁观者两组, 结果发现旁观者的存在显著降低了被试的负性情绪, 而利他性惩罚也随之减少。此外, 个体的负性情绪还会受到认知控制资源(cognitive control recourse)的影响。认知控制在解决个体的自利冲动与维护社会规范的冲突之间可能发挥着重要的作用(苏彦捷 等, 2019; Müller-Leinß et al., 2018)。然而, 认知控制依赖于有限的认知控制资源, 而认知控制资源在决策过程中会被逐渐消耗(Inzlicht & Schmeichel, 2012), 这会使个体难以控制由不公平现象引发的负性情绪, 从而实施更多的利他性惩罚(Halali et al., 2014)。

4 受控动机

4.1 促进焦点及其影响因素

自主动机从促进和预防两方面强调了惩罚者对公平的偏好: 利他性惩罚一方面是为了积极维护社会规范(促进), 另一方面则是为了消除因社会规范被破坏而产生的负性情绪(预防)。然而, 基于偏好公平的自主动机无法完全解释利他性惩罚的产生, 比如有研究者发现个体在实施利他性惩罚时对惩罚成本和收益很敏感(陈世平, 薄欣, 2016; 陈思静 等, 2020)。这意味着对外部利益的追求, 即受控动机, 同样在利他性惩罚中发挥着作用。

受控动机意味着个体是出于外部目标而做出某种行为。在利他性惩罚中, 受控动机的一个重要表现是个体对声誉(reputation)的追求。高成本信号理论(costly signaling theory)认为, 由于利他性惩罚很难为惩罚者带来直接利益, 同时这种行为通常有利于其他群体成员, 因此, 利他性惩罚本质上相当于展示了自身的某种优良特质, 如注重公平和慷慨(Nelissen, 2008)、值得信赖(Jordan, Hoffman et al., 2016)或愿意为他人付出(Jordan & Rand, 2017)等, 从而有利于惩罚者建立良好声誉, 并提高其在未来人际互动中获得他人帮助的概率。

从上述角度出发, 一个合理的推测是当人们无法通过其他亲社会行为向外界展现自己的优良特质时, 就会对他人的违规行为做出更多的惩罚。Jordan 和 Rand (2020)的研究证实了这一推测。谢东杰和苏彦捷(2019)进一步注意到, 如果旁观者可以直接推断出个人的合作意图, 那么非惩罚性合作者往往比惩罚性合作者更受青睐; 但在某些特定情境下, 惩罚可能是传达维护社会规范意愿的唯一途径, 那么个体就更愿意惩罚违规者。从这个角度来说, 利他性惩罚在一定程度上是一种基于自身利益的并由受控动机驱动的策略性行为, 即当惩罚行为对个体带来的补偿超过了成本时个体就倾向于实施惩罚。

社会地位, 即个体在他人眼中所唤起的尊崇(Torelli et al., 2019), 会显著影响惩罚动机。人们对高社会地位个体进行道德评价时往往具有“结果放大器效应”, 即认为他们的违规行为应当遭到更加严厉的惩罚(Fragale et al., 2009)。在利他性惩罚中, 低社会地位个体面对高社会地位的违规者时, 往往会通过做出更多的惩罚来体现自己的道德优越感(van Prooijen & Lam, 2007), 从而帮助其建立积极的声誉; 对高社会地位个体而言, 尽管 Blue 等(2016)的最后通牒博弈实验表明, 在面对不公平分配方案时, 他们往往倾向于实施惩罚;

但如果惩罚可能被认为是一种自私行为，他们则倾向于不实施惩罚(Vincent, 2017)，这很有可能是为了维护自己的声誉。

此外，在公开场合或个体发觉自己正在被他人观察时，往往会表现出更强的亲社会行为，以此为自己带来良好的声誉(Bereczkei et al., 2010)。即便是一些隐晦的社会线索(例如一张眼睛的图片)也会对个体的亲社会行为产生影响(Xin et al., 2016)。因此，一个合理的推测是当惩罚并非匿名时，它会具有更高的信号价值，个体也会倾向于做出更多的惩罚。然而，这里需要指出的是，旁观者的存在对个体的惩罚行为似乎产生了两种相互对抗的作用：Nelissen 和 Zeelenberg (2009)的实验显示，旁观者降低了被试的惩罚频率，因为旁观者的存在在一定程度上减轻了被试的负性情绪；而 Xin 等(2016)则指出，旁观者有利于个体做出亲社会行为，因为这有助于建立良好声誉。上述矛盾清楚地表明惩罚动机并非一个单维度概念(uni-dimensional construct)，而是一个由若干相互制约的因素所构成的体系，同一外部条件可能会对不同因素产生相反的作用，这也正是本文对动机进行二维度划分的重要原因。

4.2 预防焦点及其影响因素

Nelissen (2008)强调了利他性惩罚的动机包含趋利和避害两个面向。受控−促进焦点凸显了惩罚动机的趋利面向；与此相反，受控−预防焦点则更多地强调了对潜在损失的规避。Gardner (2019)指出，对违规行为熟视无睹有可能被其他群体成员解读为一种缺乏正义感的消极信号，因此，惩罚的实施在一定程度上是为了避免自身声誉受损。此外，Schein 和 Gray (2018)也注意到，群体成员在遵守规范的同时还会相互监督，并通过道德判断来评估观察对象的声誉，而个体为了避免被判断为道德品质不良，会尽量做出群体规范所鼓励的行为。因此，当维护公平成为群体中的规范时，个体就倾向于对破坏公平的行为施加惩罚。另一方面，利他性惩罚的预防焦点也可能是为了威慑潜在违规者(McCullough et al., 2013)。威慑理论(deterrence theory)认为，在小规模群体中，一个侵犯他人利益的人，将来也有可能侵犯你的利益(Krasnow et al., 2016)，这在一定程度上解释了利他性惩罚的受控−预防焦点的演化背景，即早期人类大部分时候都生活在规模较小的社群中(Boyd & Richerson, 2006)。尽管大规模的人际互动已成为现代社会的标志，但大失配假说(big mistake hypothesis)认为，人类祖先早期在小群体环境中演化出来的心理特质仍然在现代社会中发挥作用(Rossano, 2018)。当个体推断他们可能受到来自违规者的侵犯时，就更加倾向于实施惩罚(Delton & Krasnow, 2017)，因为不惩罚可能被解读为害怕或者顺从，从而增加了在未来遭遇不公对待的几率(Krasnow et al., 2016)。

群体内偏好(in-group favoritism)会在一定程度上影响利他性惩罚的受控−预防焦点。Delton 和 Krasnow (2017)的研究显示，当违规者是外群体成员而受害者是内群体成员时，个体实施利他性惩罚的比例显著偏高。一个合理的解释是，外群体成员对内群体成员的侵犯可能会在一定程度上被视为群体间斗争，因此，个体可能会由此推测某个内群体成员遭遇的不公对待有可能推广到包括自身在内的所有成员身上。为了避免其他成员(包括自己)遭受同样的侵犯，个体倾向于惩罚外群体违规者，以此进行威慑。另一方面，来自神经科学的研究表明，惩罚的受控−预防动机还可能受到外部压力等环境因素的影响：由于前额叶区域对惩罚是否可以阻止未来伤害很敏感(Buckholtz et al., 2008)，而压力等环境因素会扰乱前额叶的功能(Robbins & Amsten, 2009)，从而在一定程度上改变个体的惩罚模式。

5 惩罚成本对惩罚动机的影响

图 2 总结了利他性惩罚的动机分布及其各自的影响因素，从中可以看出，尽管利他动机在很大程度上驱动了利他性惩罚，但同样不应忽视以收益为目的的自利动机。从这个意义上来说，惩罚成本必然会对惩罚动机产生影响。事实上，已有学者在一定程度上验证了这一推测，如陈世平和薄欣(2016)的实验表明，随着惩罚成本的上升，个体实施利他性惩罚的比例也随之下降；此外，Aharoni 等(2019)的研究进一步显示，在成本信息并不是那么明确时，在做出惩罚决策时人们往往较少考虑成本，但当成本信息变得明确时，惩罚成本的提高显著抑制了惩罚数量；陈思静等(2020)的研究进一步发现个体不仅对成本数量敏感，对惩罚成本的形式也敏感：在控制成本数量的情况下，报复形式的成本比支付形式更易降低惩罚频

率，这可能是因为个体赋予不同的成本形式以不同的主观价值。来自神经科学的研究也证实了上述观点。研究者发现，个体往往会在惩罚带来的收益与惩罚成本之间进行权衡，一般来说，腹内侧前额皮层(vmPFC)和内侧前额皮层(mPFC)对高成本的惩罚具有更高的激活程度(Feng et al., 2016; Wang et al., 2017)，而殷西乐等(2019)发现，第三方在零成本任务和有成本任务下的惩罚差异在不同 tDCS 设置之间显著不同，这同样说明了惩罚成本在惩罚决策过程中的重要作用。

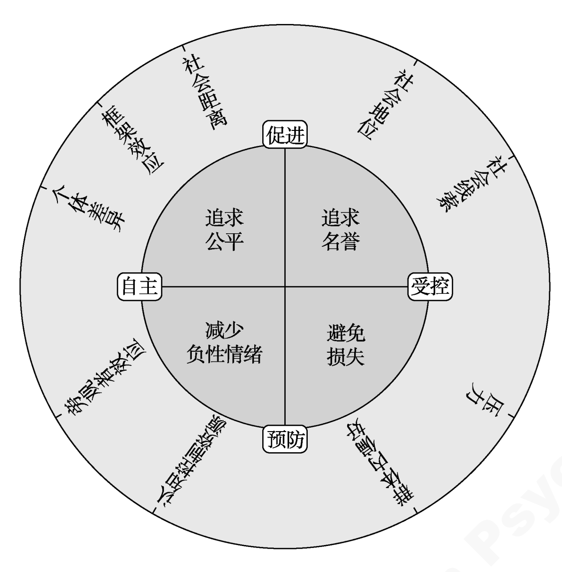


图 2 利他性惩罚动机及其影响因素的分布

遗憾的是，目前有关惩罚成本的研究通常局限在行为层面上，即相当数量的研究证实了利他性惩罚具有普通商品的属性，换言之，高成本抑制了人们“购买”惩罚的数量。然而，心理学视角的缺乏导致我们难以进一步理解成本与行为背后动机之间的关系，例如，惩罚成本的上升影响了所有的动机还是仅仅一部分动机？又如，惩罚成本对不同动机的影响是否是同方向的？举例来说，高成本的惩罚可能会增强受控−促进焦点(追求声誉)，因为按照高成本信号理论，高昂的惩罚成本会加强惩罚行为的信号传达作用，但对其他类型的动机而言，似乎不大可能存在类似作用。因此，除了行为角度以外，从动机角度出发研究惩罚成本的影响应该能极大提高对利他性惩罚的认识。

6 总结与展望

Sylwester 等(2013)指出，研究者经常在不同层面上使用“利他性惩罚”一词，这导致这一术语的含义在一定程度上变得含糊不清。从结果层面而言，大量研究表明，利他性惩罚总体上可以有效促进个体间的合作以及社会规范的维系，尽管也有学者指出利他性惩罚发挥上述作用需满足一定条件(陈思静，朱玥，2020)，如惩罚必须与相应的社会规范相结合(Bicchieri et al., 2018; Fehr & Williams, 2018)或惩罚成本必须较为低廉(Shreedhar et al., 2018)等。因此，从这个意义上来讲，利他性惩罚确实是一种利他行为，即个体通过牺牲自身利益而维护了群体规范。但从动机层面上来说，尽管无法排除他涉偏好(other-regarding preference)在利他性惩罚中的作用，然而将利他性惩罚看作全然由非策略性动机驱动似乎并不妥当。本文对利他性惩罚动机的梳理表明，以功利考虑为特征的策略性动机在利他性惩罚的实施中发挥了重要作用，尤其当惩罚动机表现出受控特征时，个体往往通过实施惩罚来为自己建立积极的声誉或避免潜在的损失。因此，从动机角度来讲，利他性惩罚似乎并不全然利他，至少在部分程度上是由自利动机驱动的。此外，对惩罚成本的讨论也在一定程度上支持了上述观点：如果利他性惩罚纯粹由利他动机驱动，那么惩罚成本在惩罚决策中似乎不应扮演重要角色，但目前的研究结论却表明，惩罚者对成本的数量和形式都十分敏感。因此，在未来研究中，如何更加准确定义利他性惩罚从而促进不同背景学者之间的交流值得我们进一步探讨。

其次，本文为强互惠(strong reciprocity)与弱互惠(weak reciprocity)之争提供了新的思路。为了解释个体在单次匿名博弈下的利他性惩罚，学者提出了强互惠理论，即由于遗传变异，某些个体具有惩罚违规者的先天倾向(Carpenter et al., 2009)。弱互惠理论家则针锋相对地指出，由于利他性惩罚的成本由惩罚者承担，但随之而来的收益却由群体共享，因此强互惠个体必然会成为搭便车对象，从而降低了其在演化中的生存概率(Dreber et al., 2008)。尽管强弱互惠之争尚无定论，但双方的共同点在于均从远因角度来解决这一争议，即利他性惩罚的演化背景，而极少考虑惩罚

者在面对具体情境时的心理机制。本文对惩罚动机的分析表明，尽管我们无法排除强互惠个体的存在，但可以肯定的是，在做出利他性惩罚的个体中至少有一部分是为了追求外在酬报(如声誉)，而非因为具有惩罚违规行为的先天倾向。有意思的是，以 Fehr 为代表的学者往往认为单次匿名博弈的设置基本排除了声誉机制对利他性惩罚的影响(Fehr & Gächter, 2002)，但 Jordan 和 Rand (2020)认为即便在单次匿名博弈中声誉动机仍在发挥作用：由于匿名并非真实生活中的常态(Dreber et al., 2008)，同时人们又经常依靠启发式来解决日常生活问题(Tversky & Kahneman, 1974)，这导致人们即使在人为的单次匿名实验设定中，仍然认为自己的行为会影响自己的声誉。综上所述，从个体心理层面来探讨强弱互惠之争应当能为我们带来新的启发。

第三，尽管本文较为全面地考察了惩罚的动机体系，但这并不意味着我们已穷尽了利他性惩罚背后所有的动机，有理由相信随着对利他性惩罚研究的不断深入，必定有更多的惩罚动机进入研究者的视野，如 Gross 等(2016)初步探讨了权力获取和利他性惩罚的关系。考虑到权力在人类社会中的重要作用，并且奖惩行为有助于个体获取权力(van Dijk et al., 2020)，一个合理的推测是，和建立积极声誉相似，权力获取同样能在一定程度上驱动利他性惩罚，即属于受控−促进焦点。此外，谢晓非等(2017)所提出的利他行为的双路径模型(dual path model)暗示，利他性惩罚的自主−促进焦点可能不仅仅意味着积极追求公平，还意味着个体主动通过实施利他性惩罚来提高自身的身心适应度。最后，对动机的梳理和分类仅仅是研究惩罚动机的第一步，更为重要的是需进一步探讨不同惩罚动机之间的关系。我们建议未来研究者关注以下问题：1)在驱动利他性惩罚时，各个动机是独立发挥作用还是互相影响？2)如果不同动机间存在交互，那么它们之间的影响是相互增强还是相互抵消？3)如果不同动机在驱动惩罚行为时存在竞争性，那么在给定情境下，哪种动机会成为焦点以及通过何种机制？目前，上述重要问题仍缺乏学界的足够关注，我们相信对上述问题的探索无疑会极大地推进对利他性惩罚的认识。

参考文献

陈世平，薄欣. (2016). 公平与惩罚价格对第三方惩罚需求的影响. *心理与行为研究*, *14*(3), 372–376.

陈思静，何铨，马剑虹. (2015). 第三方惩罚对合作行为的影响：基于社会规范激活的解释. *心理学报*, *47*(3), 389–405.

陈思静，胡华敏，杨莎莎. (2020). 支付与报复：成本形式对第三方惩罚的影响. *心理科学*, *43*(2), 416–422.

陈思静，马剑虹. (2011). 第三方惩罚与社会规范激活——社会责任感与情绪的作用. *心理科学*, *34*(3), 670–675.

陈思静，朱玥. (2020). 惩罚的另一张面孔：惩罚的负面作用及破坏性惩罚. *心理科学*, *43*(04), 911–917.

胡小勇，郭永玉. (2009). 自主-受控动机效应及应用. *心理科学进展*, *17*(1), 197–203.

李佳，蔡强，黄禄华，王念而，张玉玲. (2012). 利他惩罚的认知机制和神经生物基础. *心理科学进展*, *20*(5), 682–689.

苏彦捷，谢东杰，王笑楠. (2019). 认知控制在第三方惩罚中的作用. *心理科学进展*, *27*(8), 1331–1343.

吴燕，罗跃嘉. (2011). 利他惩罚中的结果评价——ERP 研究. *心理学报*, *43*(6), 661–673.

谢东杰，苏彦捷. (2019). 第三方惩罚的演化与认知机制. *心理科学*, *42(1)*, 216–222.

谢晓非，王逸璐，顾思义，李蔚. (2017). 利他仅仅利他吗?——进化视角的双路径模型. *心理科学进展*, *25*(9), 1441–1455.

徐杰，孙向超，董悦，汪祚军，李伟强，袁博. (2017). 人情与公正的抉择：社会距离对第三方干预的影响. *心理科学*, *40*(5), 1175–1181.

姚琦，乐国安. (2009). 动机理论的新发展：调节定向理论. *心理科学进展*, *17*(6), 1264–1273.

殷西乐，李建标，陈思宇，刘晓丽，郝洁. (2019). 第三方惩罚的神经机制：来自经颅直流电刺激的证据. *心理学报*, *51*(5), 571–583.

Aharoni, E., Kleider-Offutt, H. M., Brosnan, S. F., & Watzek, J. (2019). Justice at any cost? The impact of cost-benefit salience on criminal punishment judgments. *Behavioral Sciences & the Law*, *37*(1), 38–60.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping*, *33*(6), 1452–1469.

Bereczkei, T., Birkas, B., & Kerekes, Z. (2010). The presence of others, prosocial traits, machiavellianism. *Social Psychology*, *41*(4), 238–245.

Bicchieri, C., Dimant, E., & Xiao, E. T. (2018). *Deviant or wrong? The effects of norm information on the efficacy of punishment* (PPE Working Papers 0016). Philadelphia, PA:

Philosophy, Politics and Economics of University of Pennsylvania.

Blue, P. R., Hu, J., Wang, X., van Dijk, E., & Zhou, X. (2016). When do low status individuals accept less? The interaction between self-and other-status during resource distribution. *Frontiers in Psychology*, *7*, 1667.

Boyd, R., & Richerson, P. J. (2006). Solving the puzzle of human cooperation. In S. C. Levinson & P. Jaisson (Eds.), *Evolution and culture* (pp. 105–132). Cambridge, MA: MIT Press.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–940.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, *43*(1), 9–64.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S. H. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization*, *71*(2), 221–232.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2013). Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Social Cognitive and Affective Neuroscience*, *8*(4), 424–431.

Crowe, E., & Higgins, E. T. (1997). Regulatory focus and strategic inclinations: Promotion and prevention in decision-making. *Organizational Behavior and Human Decision Processes*, *69*(2), 117–132.

Cui, F., Wang, C., Cao, Q., & Jiao, C. (2019). Social hierarchies in third-party punishment: A behavioral and ERP study. *Biological Psychology*, *146*, 107722.

Deci, E. L., Olafsen, A. H., & Ryan, R. M. (2017). Self-determination theory in work organizations: The state of a science. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 19–43.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734–743.

Dimitroff, S. J., Harrod, E. G., Smith, K. E., Faig, K. E., Decety, J., & Norman, G. J. (2020). Third-party punishment following observed social rejection. *Emotion*, *20*(4), 713–720.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351.

Elster, J. (2006). Altruistic behavior and altruistic motivations. In S-C. Kolm & J. Mercier Ythier (Eds.), *Handbook of the economics of giving, altruism and reciprocity* (pp. 183–206). Amsterdam, Netherlands: North Holland Publishing.

Enge, S., Mothes, H., Fleischhauer, M., Reif, A., & Strobel, A. (2017). Genetic variation of dopamine and serotonin function modulates the feedback-related negativity during altruistic punishment. *Scientific Reports*, *7*, 2996.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468.

Fehr, E., & Williams, T. (2018). *Social norms, endogenous sorting and the culture of cooperation*. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677.

Fragale, A. R., Rosen, B., Xu, C., & Merideth, I. (2009). The higher they are, the harder they fall: The effects of wrongdoer status on observer punishment recommendations and intentionality attributions. *Organizational Behavior and Human Decision Processes*, *108*(1), 53–65.

Gardner, T. J. (2019). *Blaming blamers: Differential obligation to punish for third-parties compared to victims* (Unpublished doctoral dissertation). Appalachian State University

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan-Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports*, *6*, 20767.

Guala, F., & Filippin, A. (2017). The effect of group identity on distributive choice: Social preference or heuristic?. *The Economic Journal*, *127*(602), 1047–1068.

Guo, H., Song, H., Liu, Y., Xu, K., & Shen, H. (2019). Social distance modulates the process of uncertain decision-making: Evidence from event-related potentials. *Psychology Research and Behavior Management*, *12*, 701–714.

Halali, E., Bereby-Meyer, Y., & Meiran, N. (2014). Between self-interest and reciprocity: The social bright side of self-control failure. *Journal of Experimental Psychology:*

General, *143*(2), 745–754.

Harlé, K. M., & Sanfey, A. G. (2007). Incidental sadness biases social economic decisions in the Ultimatum Game. *Emotion*, *7*(4), 876–881.

Henriques, G. R. (2008). The problem of psychology and the integration of human knowledge: Contrasting Wilson's consilience with the tree of knowledge system. *Theory & Psychology*, *18*(6), 731–755.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300.

Hu, T. Y., Li, J., Jia, H., & Xie, X. (2016). Helping others, warming yourself: Altruistic behaviors increase warmth feelings of the ambient environment. *Frontiers in Psychology*, *7*, 1349.

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, *7*(5), 450–463.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*(3), 192–194.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476.

Jordan, J. J., McAuliffe, K., & Rand, D. G. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*(4), 741–763.

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, 189–202.

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57–88.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, *27*(3), 405–418.

Lerner, M. J. (1965). Evaluation of performance as a function of performer's reward and attractiveness. *Journal of Personality and Social Psychology*, *1*(4), 355–360.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15.

Müller-Leinß, J. M., Enzi, B., Flasbeck, V., & Brüne, M. (2018). Retaliation or selfishness? An rTMS investigation of the role of the dorsolateral prefrontal cortex in prosocial motives. *Social Neuroscience*, *13*(6), 701–709.

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248.

Nelissen, R. M., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*(7), 543–553.

Panchanathan, K., Frankenhuis, W. E., & Silk, J. B. (2013). The bystander effect in an N-person dictator game. *Organizational Behavior and Human Decision Processes*, *120*(2), 285–297.

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514–544.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425.

Robbins, T. W., & Arnsten, A. F. (2009). The neuropsychopharmacology of fronto-executive function: Monoaminergic modulation. *Annual Review of Neuroscience*, *32*, 267–287.

Rodrigues, J., Liesner, M., Reutter, M., Mussel, P., & Hewig, J. (2020). It's costly punishment, not altruistic: Low midfrontal theta and state anger predict punishment. *Psychophysiology*, *57*(8), e13557.

Rossano, F. (2018). Social manipulation, turn-taking and cooperation in apes: Implications for the evolution of language-based interaction in humans. *Interaction Studies*, *19*(1-2), 151–166.

Roth, G., Assor, A., Kanat-Maymon, Y., & Kaplan, H. (2007). Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning. *Journal of Educational Psychology*, *99*(4), 761–774.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Shreedhar, G., Tavoni, A., & Marchiori, C. (2018). *Monitoring and punishment networks in a common-pool resource dilemma: Experimental evidence* (GRI Working Papers 292). London, England: Grantham Research Institute on Climate and the Environment.

Strelan, P., di Fiore, C., & van Prooijen, J. W. (2017). The empowering effect of punishment on forgiveness. *European Journal of Social Psychology*, *47*(4), 472–487.

Svensson, A. (2020). Identifying motives for implementing eHealth by using Activity Theory. *Sustainability*, *12*(4), 1298.

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167–188.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*(4), 339–347.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, *2*(4), 195–206.

Torelli, C. J., Leslie, L. M., To, C., & Kim, S. (2019). Power and Status across Cultures. *Current Opinion in Psychology*, *33*, 12–17.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

van Dijk, E., de Dreu, C. K., & Gross, J. (2020). Power in economic games. *Current Opinion in Psychology*, *33*, 100–104.

van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2014). Anger and prosocial behavior. *Emotion Review*, *6*(3), 261–268.

van Prooijen, J. W., & Lam, J. (2007). Retributive justice and social categorizations: The perceived fairness of punishment depends on intergroup status. *European Journal of Social Psychology*, *37*(6), 1244–1255.

Vekaria, K. M., Brethel-Haurwitz, K. M., Cardinale, E. M., Stoycos, S. A., & Marsh, A. A. (2017). Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour*, *1*(5), 1–7.

Vincent, A. (2017). Punishment and status in collective action: How status hierarchies foster optimal punishment use. *Sociology Compass*, *11*(6), e12478.

Voelpel, S. C., Eckhoff, R. A., & Förster, J. (2008). David against Goliath? Group size and bystander effects in virtual knowledge sharing. *Human Relations*, *61*(2), 271–295.

Wang, L., Lu, X., Gu, R., Zhu, R., Xu, R., Broster, L. S., & Feng, C. (2017). Neural substrates of context-and person-dependent altruistic punishment. *Human Brain Mapping*, *38*(11), 5535–5550.

Whitson, J. A., Kim, J., Wang, C. S., Menon, T., & Webster, B. D. (2019). Regulatory focus and conspiratorial perceptions: The importance of personal control. *Personality and Social Psychology Bulletin*, *45*(1), 3–15.

Winterheld, H. A., & Simpson, J. A. (2016). Regulatory focus and the interpersonal dynamics of romantic partners' personal goal discussions. *Journal of Personality*, *84*(3), 277–290.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of United States of America*, *102*(20), 7398–7401.

Xin, Z., Liu, Y., Yang, Z., & Zhang, H. (2016). Effects of minimal social cues on trust in the investment game. *Asian Journal of Social Psychology*, *19*(3), 235–243.

Zhou, X., & Wu, Y. (2011). Sharing losses and sharing gains: Increased demand for fairness under adversity. *Journal of Experimental Social Psychology*, *47*(3), 582–588.

Motives of altruistic punishment

CHEN Sijing, YANG Shasha

(*School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou 310023, China*)

Abstract: The altruistic punishment is proposed as an important mechanism for the existence of social norms. The motives for punishing altruistically, however, are not entirely altruistic from the individual perspective. In addition to maintaining the principle of fairness, the pursuit of a good reputation, the aversion of potential losses, or the elimination of negative emotions also drive, to varying degrees, altruistic punishment. In addition, the sensitivity to the amount and form of sanction costs also shows that strategic motivations based on the cost-benefit principle play a significant role in driving altruistic punishment. Further exploration of the interaction between different motivations in the implementation of altruistic punishment is an important issue that deserves more attention in the future research.

Key words: altruistic punishment, punishment motive, fairness principle, reputation, punishment cost